Nghiên cứu độ đo khoảng cách mới cho mô hình động học gõ bàn phím trong xác thực sinh trắc học

Trần Nguyên Ngọc, Nguyễn Ngọc Hà

***Tóm tắt*— Bài báo trình bày một số vấn đề về động học gõ bàn phím (Keystroke Dynamics - KD), trong đó mô tả ngắn gọn quá trình thực hiện và một số ứng dụng của nó. Bên cạnh việc phân tích mối quan hệ giữa các độ đo khoảng cách và mô hình dữ liệu, các tác giả thực hiện các thuật toán đã có và dựa trên độ đo khoảng cách mới để cải thiện chất lượng của quá trình nhận dạng và xác thực người dùng. Các kết quả thực nghiệm dựa trên các tệp dữ liệu mẫu đã chứng tỏ độ đo khoảng cách mới có hiệu năng tốt hơn, trong đó bao gồm cả các tệp dữ liệu nhận được từ các thiết bị sử dụng màn hình cảm ứng, chẳng hạn như điện thoại thông minh.**

***Abstract*— In this paper, some problems of keystroke dynamics (KD) are presented. The substance and its use is introduced briefly. Firstly, the relationship between the distance metrics and the data model is analyzed. Then new distance based algorithm for keystroke dynamics classification is done to improve the quality of the process. Experimental results based on the sample data files proved better performance of the new distance metric. Especially including data files received from the device using the touch screen, such as smartphones.**

Từ khóa— xác thực sinh trắc học; động học gõ bàn phím; độ đo khoảng cách.

Keywords— biometric authentication; keystroke dynamics; distance metrics.

I. GIỚI THIỆU

Ngày nay, người dùng lưu trữ ngày càng nhiều dữ liệu nhạy cảm trên các thiết bị di động. Vì vậy, thực hiện các cơ chế xác thực mạnh là một yếu tố rất quan trọng. Việc phân tích các mẫu gõ bàn phím của người dùng thường sử dụng phương pháp động học gõ bàn phím. Phương pháp này rất hữu ích để nâng cao độ an toàn của cơ chế xác thực dựa trên mật khẩu. Hơn nữa, hiện nay màn hình cảm ứng của các thiết bị cho phép bổ sung các đặc tính khác nhau như áp lực lên màn hình và diện tích tiếp xúc ngón tay vào các đặc tính dựa trên thời gian... để dùng cho mô hình động học gõ bàn phím. Đã có những nghiên cứu kiểm chứng được hiệu quả của việc bổ sung đặc tính màn hình cảm ứng đối với việc nhận dạng và xác thực thông qua bộ dữ liệu của người dùng. Kết quả cho thấy rằng các thuộc tính bổ sung này nâng cao tính chính xác của cả hai quá trình. Trong nội dung nghiên cứu này, chúng tôi trình bày các kết quả thử nghiệm áp dụng độ đo mới đối với hai tập dữ liệu là CMU [4] và nhóm Margit Antall [1]. Kết quả thu được cho thấy, độ đo mới này giúp nâng cao hơn nữa độ chính xác của quá trình xác thực bằng việc lựa chọn giá trị vectơ trung tâm (medium vector).

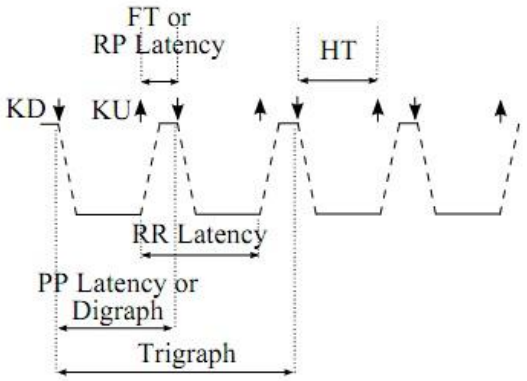
Bố cục của bài báo gồm 5 Mục. Sau Mục Giới thiệu, trong Mục II trình bày tóm tắt lĩnh vực nghiên cứu KD, xem xét một vài nghiên cứu đã được thực hiện trên các thiết bị di động có sử dụng màn hình cảm ứng. Mục III chúng tôi tham khảo hai bộ dữ liệu điển hình CMU [4] và nhóm Margit Antall [1] để đề xuất độ đo mới và trong Mục IV thực hiện thử nghiệm và đánh giá hiệu quả của độ đo mới. Mục cuối cùng, trình bày một số kết luận và hướng nghiên cứu tương lai.

II. ĐỘNG HỌC GÕ BÀN PHÍM

Động học gõ bàn phím là một lĩnh vực đã được nghiên cứu nhiều, do việc triển khai dễ dàng và có chi phí thấp. Trái ngược với các phương pháp sinh trắc học khác, phương pháp này không yêu cầu bất kỳ thiết bị phần cứng chuyên dụng nào khác. Việc lấy mẫu động học gõ bàn phím được thực hiện chỉ bằng phần mềm chạy nền, điều này đảm bảo tính khả thi và không gây ảnh hưởng đến người dùng. KD có thể sử dụng cho cả hai trường hợp là xác thực theo thời điểm và xác thực liên tục. Tuy nhiên, so với các phương pháp xác thực khác thì phương pháp sinh trắc học này cũng có nhược điểm là độ chính xác thấp hơn.

Dữ liệu dùng trong các báo cáo nghiên cứu KD được thu thập từ nhiều thiết bị đầu vào khác nhau, từ bàn phím thường đến những bàn phím có cảm ứng áp lực. Phổ biến nhất là sử dụng đặc tính dựa trên thời gian, đó là thời gian Dwell và thời gian Flight. Dwell là khoảng thời gian tính từ lúc nhấn một phím đến lúc nhả phím đó ra (đôi khi được gọi là thời gian giữ), Flight là khoảng thời gian từ lúc nhả phím này đến lúc nhấn một phím tiếp theo.

Một số trường hợp có thể có ba hoặc nhiều hơn các sự kiện thời gian liên tiếp liên quan đến gõ phím được sử dụng như các đặc tính N-graphs (Hình 1) [9,10].



Hình 1. Đặc tính của 3 – graphs

Trong đó:

KU: thời điểm nhả phím (Key up/release);

KD: thời điểm nhấn phím (Key down/release);

FT: là khoảng thời gian từ lúc nhả phím này đến lúc ấn phím tiếp theo (Fight time);

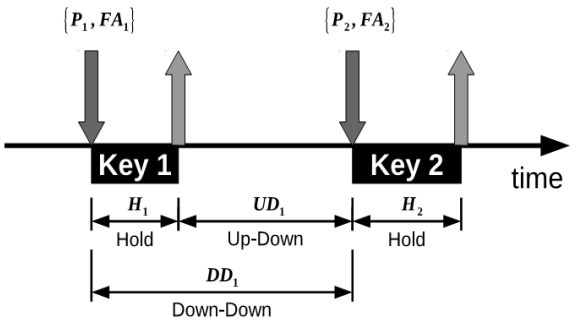
HT: khoảng thời gian giữ phím (Hold/dwell time);

PP: là khoảng thời gian từ lúc ấn phím này đến lúc nhấn phím tiếp theo (Press-press);

RR: khoảng thời gian từ lúc nhả phím này đến lúc nhả phím tiếp theo (Release-release);

RP: khoảng thời gian từ lúc nhả phím đến lúc nhấn phím tiếp theo (Release-press).

Hầu hết các đề xuất nhận dạng mẫu người dùng đều được thực hiện bằng nhận dạng động học gõ bàn phím, kể cả các phương pháp tiếp cận thống kê và học máy. Trong phần lớn các bài báo [2, 4] chỉ sử dụng các đặc tính gõ hai phím liên tiếp nhau, như trong Hình 2.



Hình 2. Đặc tính 2-graphs từ thiết bị di động [1]

Trong đó:

H: thời gian giữ phím (Hold time);

UD: khoảng thời gian giữa hai phím (Up-Down time);

DD: khoảng thời gian chuyển phím (Down-Down time);

P: áp lực ngón tay (Pressure);

FA: diện tích tiếp xúc ngón tay (Finger Area).

Phương pháp xác thực đơn giản là xây dựng mẫu tham chiếu cho người sử dụng và tính khoảng cách (độ xa) giữa mẫu gõ hiện thời với mẫu gõ tham chiếu tương ứng. Phương pháp này gọi là so khớp mẫu và có thể được phối hợp với các phép đo khác, như phép đo đơn giản Euclide và phép đo phức tạp hơn là Mahalanobis [3]. Trong đó mạng Nơron và máy vectơ (SVM) là hai hệ thống hỗ trợ phương pháp này là tốt nhất. Ngoài ra, còn một số bộ phát hiện sai khác nữa bao gồm: Nearest-neighbor, Fuzzy-logic, Outlier-counting (z-score), Fc-means.

Hệ thống sinh trắc học thường bao gồm hai chức năng riêng biệt là xác thực và nhận dạng [1]. Xác thực là một bài toán xác định nhị phân, mà hệ thống chấp nhận hoặc từ chối danh tính đã được đăng ký bởi người dùng. Và nhận dạng (còn được gọi là sự công nhận) là một bài toán phân loại: Hệ thống phân loại các mẫu dữ liệu nhập vào thành một trong *N* lớp đã được biết trước.

Chất lượng của hệ thống sinh trắc học thường được đặc trưng bởi ba loại lỗi: tỷ lệ chấp nhận sai (FAR), tỷ lệ từ chối sai (FRR) và tỷ lệ mà tại đó FAR là bằng với FRR (EER). Với FAR, một hệ thống sinh trắc học chấp nhận một mẫu thuộc về danh tính đã được đăng ký lại thuộc về một mẫu của kẻ mạo danh. Với FRR, một hệ thống sinh trắc học từ chối không đúng một mẫu cung cấp bởi người sử dụng đã đăng ký.

III. ĐỀ XUẤT ĐỘ ĐO MỚI

Mô hình KD là một cách thức để nhận được đại diện cho bộ dữ liệu KD của người dùng. Phần lớn các công trình [5, 3, 6] tập trung vào việc lựa chọn độ đo khoảng cách và đã chọn véctơ trung bình (mean vector) [7] như một mô hình mặc định. Tuy nhiên véctơ trung bình không thực sự đảm bảo khoảng cách tối ưu của đám mây giữa các điểm KD. Ví dụ sau đây sẽ minh họa tình huống này.

Ví dụ: Cho một tập gồm 3 véctơ $X = \{x_1 = (7,6,9), x_2 = (2,3,5), x_3 = (9,6,7)\}$.

Véctơ trung bình là $\bar{x} = (6,5,7)$, véctơ trung tâm (median vector) là $m = (7,6,7)$ được hình thành bởi các giá trị trung tâm. Các khoảng cách Manhattan giữa mỗi véctơ và X được tính toán như sau:

$$dist(X,\bar{x}) = \left(|7-6|+|6-5|+|9-7|\right)+$$
$$\left(|2-6|+|3-5|+|5-7|\right)+\left(|9-6|+|6-5|+|7-7|\right)=16$$
$$dist(m,X) = \left(|7-7|+|6-6|+|7-9|\right)+$$
$$\left(|7-2|+|6-6|+|7-9|\right)+\left(|7-9|+|6-6|+|7-7|\right)=9$$

Do đó, $dist(X,\bar{x}) > dist(X,m)$.

Như vậy, mô hình KD có thể được xác định theo các cách tiếp cận khác nhau. Phương pháp Weiszfeld nổi tiếng [8] dùng để giải bài toán tìm vị trí tiện ích đã chỉ ra rằng: Cho $X := \left\{ x_i : i \in \overline{1:N} \right\}$ là một tập N điểm dữ liệu trong R^n, để tìm ra một điểm $c \in \mathrm{R}^n$ làm cực tiểu với tổng các khoảng cách:

$$dist(X,c) = \min_{c \in \mathrm{R}^n} \sum_{i=1}^{N} d(x_i,c), \qquad (1)$$

trong đó: $d(p,q) = \|p-q\|$ ký hiệu khoảng cách Euclid của hai véctơ $p,q \in \mathrm{R}^n$, thì ta cần phải sử dụng phép tính lặp Weiszfeld. Gradient của $d(X,x)$ sẽ không xác định nếu c trùng với một trong các điểm dữ liệu x_i. Đối với $c \notin X$, ta có:

$$\nabla dist(X,c) = -\sum_{i=1}^{N} \frac{x_i - c}{\|x_i - c\|}. \qquad (2)$$

Trung tâm tối ưu được ký hiệu là c^*, nếu không có trong X sẽ được đặc trưng bởi $\nabla dist(X,c^*) = 0$, khi đó biểu diễn nó như một tập lồi của các điểm x_i, $c^* = \sum_{i=1}^{N} \lambda_i x_i$, với các trọng số:

$$\lambda_i = \frac{1/\|x_i - c\|}{\sum_{k=1}^{N} 1/\|x_k - c\|} \text{ phụ thuộc vào } c^*.$$

Kết quả quay vòng này dẫn tới phép lặp Weiszfeld như sau

$$c_+ := T(c) \qquad (3)$$

trong đó: c_+ là trung tâm được cập nhật, c là trung tâm hiện thời, còn

$$T(c) := \begin{cases} \dfrac{\sum_{i=1}^{N} x_i / \|x_i - c\|}{\sum_{k=1}^{N} 1/\|x_k - c\|} & \text{khi } c \notin X \\ c, & \text{khi } c \in X \end{cases}. \qquad (4)$$

Hơn nữa, nếu ta cũng sử dụng lược đồ tương tự này cho khoảng cách Mahattan, khi đó:

$$d(p,q) = \|p-q\|_{L1} = \sum_{i=1}^{N} |p^i - q^i|; p,q \notin \mathrm{R},$$

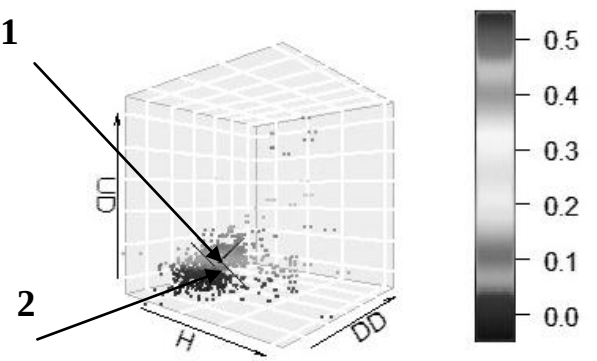
gradient của $dist(X,c)$ được biểu diễn như sau:

$$\nabla dist(X,c) = \sum_{i=1}^{N} sign(x_i - c). \qquad (5)$$

Trong trường hợp này, c^* tối ưu là véctơ trung tâm chứ không phải là véctơ trung bình của X.

Phần trình bày ở trên chứng tỏ rằng nếu sự kiện là mô hình KD (trung tâm của đám mây KD) thì phải được nhận diện với độ đo khoảng cách tương ứng.

Quá trình khảo sát tập dữ liệu thu thập thực tế về KD cho thấy, việc lấy giá trị trung bình (mean) của các đặc trưng có thể không phản ánh tốt bản chất tập dữ liệu (Hình 3), vì có nhiều dữ liệu trong đó phân bổ khác xa so với phần còn lại (một số trường hợp người dùng gõ sai nhịp điệu thông thường). Do vậy trong nghiên cứu này, thay vì chọn các giá trị trung bình (Mean) cho x_i, chúng tôi đề xuất lựa chọn giá trị nằm giữa tập dữ liệu huấn luyện (Median).



Hình 3. Phân bổ dữ liệu KD trên đồ thị 3D
1-Mean (trung bình); 2-Median (trung tâm)

Theo đó, bộ các trọng số a_i thực chất là xét tới ảnh hưởng của các đặc trưng thành phần trong véctơ $\bar{x}$. Với những đặc trưng có phân bố "loãng" độ lệch chuẩn lớn thì độ tin cậy kém hơn, những đặc trưng tương đối "chụm" độ tin cậy sẽ

cao hơn. Tuy nhiên, mức độ ảnh hưởng này nếu thuần túy lựa chọn cách tổ hợp tuyến tính, thì không làm nổi bật nhanh những đặc trưng tốt, có độ chụm cao, do các đặc trưng có hình dạng phân bố khác nhau (Hình 3).

Độ đo khoảng cách là một vấn đề quan trọng trong các thuật toán phân loại. Độ đo khoảng cách thường được sử dụng là Euclide, Manhattan,… đều giả thiết rằng mỗi đặc trưng của điểm dữ liệu có tầm quan trọng như nhau và độc lập với các đặc trưng khác. Giả thiết này không phải bao giờ cũng được thỏa mãn trong các tình huống thực tế, đặc biệt khi xử lý dữ liệu có số chiều lớn mà trong đó một số đặc trưng của điểm dữ liệu có thể không liên quan chặt chẽ với đặc trưng được xét đến.

Ngược lại, một độ đo khoảng cách với chất lượng tốt cần nhận diện được các đặc trưng quan trọng và phân biệt được các đặc trưng nào có liên quan và không liên quan. Do đó, việc cung cấp một độ đo khoảng cách là một vấn đề đặc biệt quan trọng và quyết định sự thành công hay thất bại của thuật toán học hoặc của hệ thống được phát triển.

Ý tưởng cơ bản của bài báo này là lựa chọn và biến đổi các đặc trưng dữ liệu thành một không gian đặc trưng mới được chuẩn hóa hơn. Sau đó thiết kế một độ đo khoảng cách mới, trong đó có tính đến những mối quan hệ không mong muốn giữa các đặc trưng và làm giảm ảnh hưởng của các điểm bất thường nhằm cải thiện hiệu năng. Về mặt hình thức, ở giai đoạn huấn luyện chúng tôi áp dụng ý tưởng tương tự như trong [3] để lọc và xác định tỷ lệ xích dữ liệu huấn luyện. Trước hết, các véctơ với độ lệch cao sẽ bị loại ra khỏi bộ dữ liệu huấn luyện bằng cách sử dụng khoảng cách Mahattan:

$$X^* := \left\{ x : x \in X, \left\| x - m \right\|_{L_1} \le \frac{\sum_{k=1}^{N} \left\| x_k - m \right\|_{L1}}{N} \right\} \quad (6)$$

trong đó: $m = median(X)$.

Sau khi nhận được bộ dữ liệu huấn luyện có lọc X^*, từng đặc trưng trong mỗi véctơ $x_i \in X^*, i = \overline{1...M}$ được xác định tỷ lệ xích bằng độ lệch tuyệt đối trung bình $\sigma = \left(\sigma^1, \sigma^2, ..., \sigma^n\right)$:

$$x_i = \left(x_i^1, x_i^2, ..., x_i^n\right) := \left(\frac{x_i^1}{\sigma^1}, \frac{x_i^2}{\sigma^2}, ..., \frac{x_i^n}{\sigma^n} \right), \quad (7)$$

trong đó:

$$\sigma^k = \frac{\sum_{j=1}^{M} \left| x_j^k - \frac{\sum_{l=1}^{M} x_l^k}{M} \right|}{M}. \quad (8)$$

Độ đo khoảng cách mới được đề xuất sau đây sẽ đánh giá thích hợp khoảng cách bằng cách gán các hệ số chỉ mức độ quan trọng khác nhau cho các đặc trưng của các điểm dữ liệu. Nó được xác định bằng một hàm phi tuyến có dạng:

$$dist(\bar{x}, \bar{y}) = \sum_{i=1}^{N} \ln(1 + |x_i - y_i| / a_i) \quad (9)$$

Từ đó thấy rằng hàm logarithm cho khoảng cách $dist(\bar{x}, \bar{y})$ bền vững hơn với các thay đổi lớn (cho mục đích từ chối kẻ mạo danh) và hàm này nhạy hơn với các thay đổi nhỏ (cho mục đích chấp nhận người dùng hợp pháp). Mặt khác, ta có thể dễ dàng tính được véctơ gradient và mô hình tối ưu (trung tâm):

$$\nabla dist(X, c) = \nabla \sum_{i=1}^{M} d_N(x_i, c) = 2 \sum_{i=1}^{M} \frac{x_i - c}{1 + (x_i - c)^2} \quad (10)$$

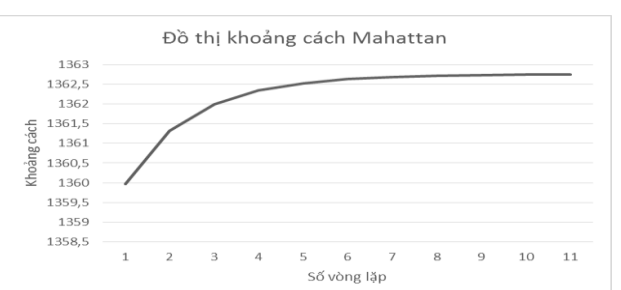
Chú ý rằng, các toán tử trên tác động lên từng phần tử. Khi cho gradient (10) bằng 0 ta có:

$$\sum_{i=1}^{M} \frac{x_i}{1 + (x_i - c)^2} = c \sum_{i=1}^{M} \frac{1}{1 + (x_i - c)^2} \quad (11)$$

Phương trình (11) bao gồm ánh xạ:

$$c_+ := \frac{\sum_{i=1}^{M} \frac{x_i}{1 + (x_i - c)^2}}{\sum_{i=1}^{M} \frac{1}{1 + (x_i - c)^2}} \quad (12)$$

Cuối cùng, sự hội tụ của phép lặp (12) có thể được thiết lập như trong thuật toán Weiszfeld [8]. Sự hội tụ của quá trình huấn luyện được minh họa bằng thực nghiệm số như Hình 4.



Hình 4. Đồ thị sự hội tụ của khoảng cách bằng phép lặp Weiszfeld

Quá trình huấn luyện có thể được mô tả trong Thuật toán 1 với đầu ra là mẫu (template) δ. Mẫu δ này được dùng để xây dựng mẫu tham chiếu cho người sử dụng liên quan và tính toán khoảng cách giữa mẫu gõ hiện thời với mẫu tham chiếu ở giai đoạn xác thực người dùng như được mô tả trong Thuật toán 2.

Thuật toán 1: Tính toán mô hình KD

INPUTS: $X \coloneqq \left\{ x_i : i \in \overline{1:N} \right\}$ tập các véctơ đặc trưng; ngưỡng ε.

Xử lý sơ bộ: Lọc X khi sử dụng (6); tính toán véctơ σ khi sử dụng (8) và tập dữ liệu mới X^* với M véctơ khi sử dụng (7).

Khởi tạo: gán mô hình c bằng véctơ trung tâm của X^*

Quá trình lặp:

1. Tính khoảng cách :

$$dist(X^*, c) = \sum_{i=1}^{M} d_N(x_i, c)$$

2. Cập nhật mô hình c_+ khi sử dụng (12)

3.Tính toán khoảng cách mới :

$$dist(X^*, c_+) = \sum_{i=1}^{M} d_N(x_i, c_+)$$

4. **if** $\left| dist(X^*, c) - dist(X^*, c_+) \right| < \varepsilon$ **then**

5. **return** template $\delta = \{c, \delta\}$

6. **end if**

Thuật toán 2: Kiểm chứng mô hình KD

INPUTS: $X \coloneqq \left\{ x_i : i \in \overline{1:N} \right\}$ tập các véctơ đặc trưng KD; template $\delta = \{c, \delta\}$; ngưỡng θ.

Xử lý sơ bộ: xác định tỷ xích X khi sử dụng (7) với véctơ σ trong template δ.

Kiểm tra:

1. Tính khoảng cách:

$$dist(X^*, c) = \sum_{i=1}^{M} d_N(x_i, c)$$

2. **if** $dist(X^*, c) < \theta$ **then**

3. **return** TRUE

4. **else**

5. **return** FALSE

6. **end if**

IV. CÁC KẾT QUẢ THỰC NGHIỆM

Năm 2009, Killorhy và Maxion đã thu thập và công bố bộ dữ liệu băng thử động học gõ bàn phím (bộ dữ liệu CMU) [4] gồm 51 chủ thể với 400 KD được thu thập cho mỗi người. Họ đã đánh giá 14 thuật toán động học gõ bàn phím khả dụng dựa trên bộ dữ liệu này. Những độ đo khoảng cách khác nhau đã được sử dụng, gồm có khoảng cách Euclide và Mahattan.

Năm 2014, Margit ANTAL, László Zsolt SZABÓ, Izabella LASZLO đã thu thập và công bố bộ dữ liệu thu thập từ bàn phím trên các thiết bị di động nền Android [2] có màn hình cảm ứng gồm 42 chủ thể với 51 KD được thu thập cho mỗi người. Họ đã chứng minh bằng thực nghiệm rằng các thuộc tính dựa trên màn hình cảm ứng cải thiện đáng kể phương pháp động học gõ bàn phím trong việc phân loại và xác thực. Trong các phép đo nhận dạng, việc bổ sung các đặc tính từ màn hình cảm ứng vào tập đặc tính mặc định đã làm tăng hơn 10% độ chính xác cho việc phân lớp. Việc cải thiện này khó giải thích bằng lý thuyết trong trường hợp các phép đo xác thực vì tỷ lệ EER chỉ giảm khoảng 2,4% (trong phép đo Manhattan).

A. Kết quả xác thực bằng các phép đo tập dữ liệu của nhóm Killourhy & Maxion

Các phép đo để xác thực được thực hiện bằng cách sử dụng tập lệnh ngôn ngữ R được cung cấp bởi Killourhy & Maxion [3]. Tập lệnh này cung cấp phép tính tỷ lệ EER cho ba bộ phát hiện sự bất thường dựa trên các phép đo Euclide, Manhattan, và Mahalanobis.

Tên tập dữ liệu datafile <- 'DSL-StrongPasswordData.txt';

Chuỗi ký tự được nhập: .tie5Roanl;

Số thuộc tính cho một lần nhập: 31 thuộc tính (time based);

Số người khảo sát: 51 người;

Số lần nhập: 400 lần;

Tổng số 8 phiên (50 lần/ phiên).

Sau đây là các bảng kết quả phép đo:

BẢNG 1. KẾT QUẢ ĐO VỚI SỐ LẦN HUẤN LUYỆN LÀ 200 LẦN

Detector	**eer.mean (Killorhy và Maxion)**	**eer.mean**	**eer.sd**
Euclide	0,171	0,171	0,095
Manhattan	0,153	0,153	0,092
Mahalanobis	0,110	0,110	0,065
ManhattanScaledH		0,062	0,056

BẢNG 2. KẾT QUẢ ĐO VỚI SỐ LẦN HUẤN LUYỆN LÀ 250 LẦN

Detector	**eer.mean**	**eer.sd**
Euclide	0,159	0,087
Manhattan	0,142	0,089
Mahalanobis	0,106	0,069
ManhattanScaledH	0,059	0,055

BẢNG 3. KẾT QUẢ ĐO VỚI SỐ LẦN HUẤN LUYỆN LÀ 300 LẦN

Detector	**eer.mean**	**eer.sd**
Euclide	0,151	0,087
Manhattan	0,133	0,090
Mahalanobis	0,104	0,072
ManhattanScaledH	0,057	0,061

BẢNG 4. KẾT QUẢ ĐO VỚI SỐ LẦN HUẤN LUYỆN LÀ 350 LẦN

Detector	**eer.mean**	**eer.sd**
Euclide	0,145	0,082
Manhattan	0,124	0,086
Mahalanobis	0,098	0,080
ManhattanScaledH	0,053	0,064

Kết quả đo ở Bảng 1 cho thấy rằng với phép đo mới này giảm ERR xuống còn 0,062 (tức là tăng hiệu quả được 43,63% so với phép đo tốt nhất hiện nay là Mahalanobis với ERR = 0,110). Nếu tiếp tục tăng số mẫu huấn luyện lên 350 lần thì tăng được hiệu quả lên đáng kể (Bảng 4).

B. Kết quả xác thực bằng các phép đo tập dữ liệu của nhóm Margit Antal

Các phép đo để xác thực được thực hiện bằng cách sử dụng tập lệnh ngôn ngữ R được cung cấp bởi Killourhy & Maxion [3]. Tập lệnh này cung cấp phép tính tỷ lệ EER cho ba bộ phát hiện bất thường dựa trên các phép đo Euclide, Manhattan, và Mahalanobis. Dữ liệu được chuẩn hóa và chia thành 3 phần bằng nhau, mỗi phần chứa 17 mật khẩu của mỗi người dùng. Hai phần đầu tiên được sử dụng để tạo mẫu người dùng và phần còn lại để kiểm tra FRR. Năm mẫu mật khẩu đầu tiên từ dữ liệu của mỗi người, trừ mẫu người kiểm tra, được dùng để kiểm tra FAR.

Tên tập dữ liệu là:

datafile <- 'keystroke_normalized.arff';

Chuỗi ký tự được nhập: .tie5Roanl;

Số thuộc tính cho một lần nhập: 71 thuộc tính (time based + touchscreen);

Số người khảo sát: 42 người;

Số lần nhập: 51 lần;

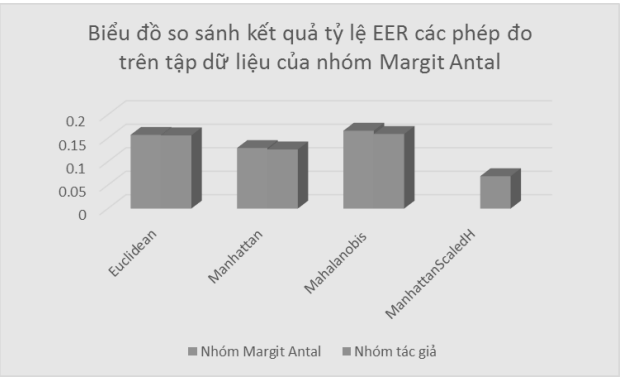
Kết quả chia làm 3 phần bằng nhau, 2 phần đầu làm dữ liệu huấn luyện, phần còn lại làm mẫu kiểm tra FFR. 5 mẫu đầu tiên (trừ mẫu kiểm tra) các tập dữ liệu khác được dùng kiểm tra dữ liệu mạo danh FAR), tổng số có 2 phiên.

Bảng kết quả như sau:

BẢNG 5. KẾT QUẢ EER CỦA PHÉP ĐO MỚI TRÊN TẬP DỮ LIỆU NHÓM MARGIT ANTAL.

Detector	**eer.mean (Margit Antal et al)**	**eer.mean**	**eer.sd**
Euclide	0,157	0,156	0,154
Manhattan	0,129	0,126	0,129
Mahalanobis	0,166	0,159	0,099
Manhattan ScaledH		0,069	0,071

Ta có biểu đồ so sánh kết quả EER như sau:



Hình 5. Biểu đồ so sánh kết quả EER

Từ Hình 5 ta thấy, với phép đo mới này, tỷ lệ EER đã tăng khoảng 45%.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã nghiên cứu mối quan hệ giữa các độ đo khoảng cách với mô hình dữ liệu động học gõ bàn phím và đề xuất một độ đo khoảng cách mới để áp dụng vào trong các thuật toán nhận dạng và xác thực. Chúng tôi đã thực hiện một loạt các thực nghiệm khi sử dụng các độ đo khoảng cách khác nhau. Các kết quả thực nghiệm cho thấy, với độ đo mới này có thể giảm ERR xuống còn 0,069.

Nghiên cứu tiếp theo sẽ tập trung vào việc giảm số mẫu thuận lợi (lượng dữ liệu được thu thập cho người dùng). Việc thử nghiệm độ đo mới sẽ được thực hiện trong môi trường sử dụng mật khẩu của người dùng và các thiết bị di động trong

thực tế, đồng thời tiến hành thử nghiệm trên các thiết bị không sử dụng hệ điều hành Anroid, như Iphone. Sau đó chúng tôi đưa thêm đánh giá độ entropy cho mỗi mật khẩu người dùng thực tế hoặc bổ sung thêm đặc tính mới hoặc tiếp tục cải tiến độ đo mới này. Hơn nữa chúng tôi cũng sẽ tập trung vào những điểm khác biệt giữa động học gõ bàn phím tĩnh và liên tục về phương diện tạo ra mô hình dữ liệu và thuật toán phân loại.

TÀI LIỆU THAM KHẢO

[1]. M. Antal, L. Z. Szabo, and I. Laszlo, "Keystroke dynamics on android platform", Procedia Technology 19, pp. 820-826, 2015.

[2]. M. Antal, L. Z. Szabo, and Laszlo, "Keystroke Dynamics - Data Set".

[3]. Kevin S. Killourhy and Roy A. Maxion, "Comparing Anomaly Detectors for Keystroke Dynamics", in Proceedings of the 39th Annual International Conference on Dependable Systems and Networks (DSN-2009), pp. 125-134, Estoril, Lisbon, Portugal, June 29-July 2, 2009. IEEE Computer Society Press, Los Alamitos, California, 2009.

[4]. R. M. Kevin Killourhy, "Keystroke Dynamics - Benchmark Data Set".

[5]. R. Giot, M. El-Abed, and C. Rosenberger, "Keystroke dynamics authentication. Biometrics", chapitre-8, 2011.

[6]. Y. Zhong, Y. Deng, and A. K. Jain. "Keystroke dynamics for user authentication". In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pp. 117-123, 2012.

[7]. M. Ferrer, E. Valveny, F. Serratosa, I. Bardaji, and H. Bunke, "Graph-based k-means clustering: A comparison of the set median versus the generalized median graph", In Computer Analysis of Images and Patterns.

[8]. A. Beck, M. Teboulle, and Z. Chikishev, "Iterative minimization schemes for solving the single source localization problem", SIAM Journal on Optimization, 19(3), pp. 1397-1416, 2008, Springer, pp. 42-350, 2009.

[9]. Trojahn M, Arndt F, Ortmeier F, "Authentication with Keystroke Dynamics on Touchscreen Keypads - Effect of different N-Graph Combinations", In: MOBILITY 2013, The Third International Conference on Mobile Services, Resources and Users, pp. 114-119, 2013.

[10]. F. Bergadano, D. Gunetti, and C. Picardi, "User authentication through keystroke dynamics". ACM Transactions on Information and System Security, 5(4): pp. 367-397, 2002.

SƠ LƯỢC VỀ TÁC GIẢ

TS. Trần Nguyên Ngọc

Đơn vị công tác: Khoa Công nghệ thông tin – Học viện Kỹ thuật Quân sự.

Email: tonono79@yahoo.com

Quá trình đào tạo: Nhận bằng Kỹ sư chuyên ngành Điều khiển và Tin học trong các hệ thống kỹ thuật tại Học viện Kỹ thuật quân sự năm 2005. Nhận bằng Tiến sĩ chuyên ngành Phân tích hệ thống, điều khiển và xử lý thông tin, năm 2007.

Hướng nghiên cứu hiện nay: An toàn thông tin.

ThS. Nguyễn Ngọc Hà

Đơn vị công tác: Viện Kỹ thuật cơ giới quân sự - Tổng cục Kỹ thuật.

Email: hayeuot@yahoo.com

Quá trình đào tạo: Nhận bằng Kỹ sư tại Đại học Bách khoa Hà Nội năm 2004, nhận bằng Thạc sĩ năm Học viện Kỹ thuật Quân sự chuyên ngành Khoa học máy tính năm 2013.

Hướng nghiên cứu hiện nay: An toàn thông tin.